Published in: Granger, Sylviane/Paquot, Magali (Hrsg.): eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009. - Louvain: UCL Presses Universitaires, 2010. p. 157-163. (Cahiers du Cental, no. 7)

On the combination of automated information and lexicographically interpreted information in two German online dictionaries

Annette Klosa[1]
Institut für Deutsche Sprache, Mannheim

Abstract

This paper discusses the advantages and disadvantages of the combination of automated information and lexicographically interpreted information in online dictionaries, namely *elexiko*, a hypertext dictionary and lexical data information system of contemporary German (http://www.owid.de/elexiko_/index.html), and DWDS, a digital dictionary of 20th century German (http://www.dwds.de). Examples of automatically derived information (*e.g.* automatically extracted citations from the underlying corpus, lists on paradigmatic relations) and lexicographically compiled information (*e.g.* information on paradigmatic partners) are provided and evaluated, reflecting on the need to develop guidelines as to how computerised information and lexicographically interpreted information may be combined profitably in online reference works.

Keywords: online dictionary, German, automated information, lexicographically interpreted information.

1. Introduction

The availability of large electronic corpora has changed the work of lexicographers in more ways than one. By applying corpus-driven and corpus-based approaches (*cf.* Tognini-Bonelli 2001), thus exploiting large electronic text corpora, modern monolingual and bilingual dictionaries are developed. The ways in which this has affected the process of compiling a dictionary have been described for specific dictionary projects (*cf.* Baugh, Harley and Jellis 1996; Sinclair 1987) and reflected on in a broader approach in various publications (*cf.* Klosa 2007; Teubert 1999; Euralex Bibliography of Lexicography: http://euralex.pbwiki.com/Corpus+Lexicography, and OBELEX: http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht; keyword=corpus-based lexicography).

But data from large corpora is not only interpreted by lexicographers: it is also the basis for computer linguistic tools and their applications such as collocation analysis. Computational linguists have developed procedures often based on statistical methods

[1] klosa@ids-mannheim.de

and designed to calculate frequency and significance, or for part-of-speech tagging. They have contributed, for example, to the development of lexical-semantic resources (*e.g.* ontologies, *cf.* Mönnich and Kühnberger 2008), or automatic sense disambiguation (*cf.* Agirre and Edmonds 2006). Computational lexicographers, in particular, have been concerned with questions of how to build a lexicon (*cf.* Boguraev 1993).

Within this context, the idea of automating and, thus, possibly accelerating the compilation of dictionaries has emerged. In the past, working on a dictionary was an exclusively human task. Nowadays, it is a combination of applying computer and corpus tools together with the lexicographer's linguistic competence. In printed dictionaries, this mainly leads to an improvement in the quality of lexicographic information but not necessarily to new types of lexicographic information. In electronic dictionaries, this can (and maybe even should) be different. Electronic dictionaries and online dictionaries, in particular, are not subjected to limitations of space. As well as the classic inventory of grammatical, morphological, orthographic, semantic, and pragmatic information, electronic dictionaries are able to offer more detailed information and new types of linguistic detail. They may also present data in non–traditional ways, for example in graphs or by using other media, such as video or audio files.

In this situation, lexicographers have to assess what computational linguistics can offer. They have to decide which information in the dictionary must still be manually compiled (*e.g.* the paraphrase of the headword) and which information might be automatically extracted (*e.g.* information on part of speech or inflection). The advantages and disadvantages of the combination of automated information and lexicographically interpreted information in online dictionaries are discussed here by looking at two German dictionaries: *elexiko*, a hypertext dictionary and lexical data information system of contemporary German (http://www.owid.de/elexiko_/index.html), and DWDS, a digital dictionary of 20th century German (http://www.dwds.de).

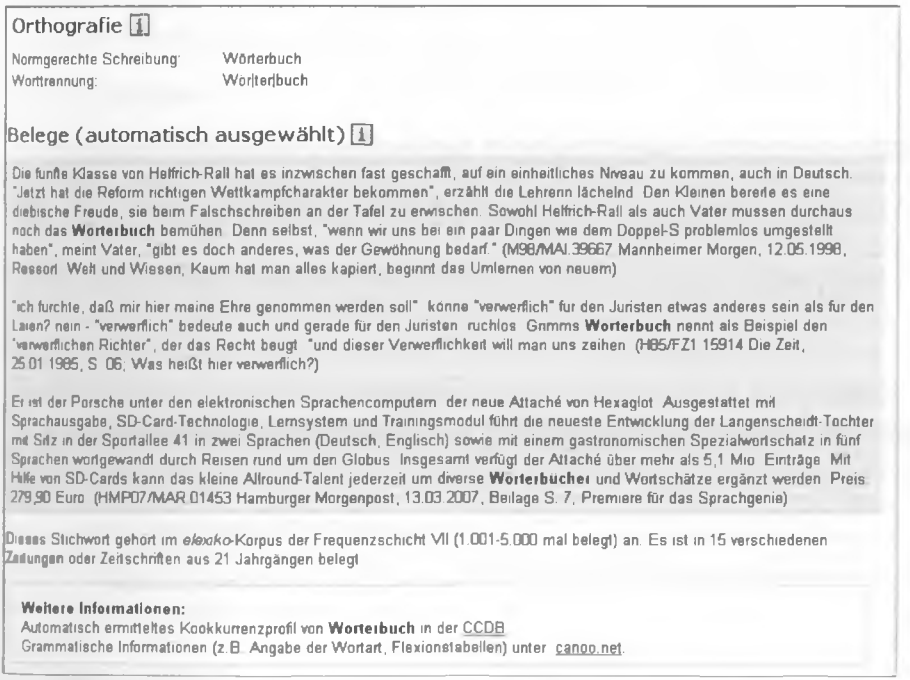
2. *elexiko*

elexiko is a lexicological-lexicographic project based at the Institute for the German Language (IDS) at Mannheim (*cf.* Haß 2005; Klosa, *et al.* 2006; Storjohann 2005b). The aim of this project is to compile a reference work, specifically designed for online publication, that explains and documents contemporary German. The primary and exclusive basis for lexicographic interpretation is an extensive German corpus. Filling *elexiko* in modules is (besides the corpus-based approach) one of the two main lexicographic methods for the dictionary. *elexiko* is compiled not in alphabetical order but by analysing the semantic, syntactic, or morphological features of the lexicon systematically in batches. Thus, a complete word class, an entire word family, or a semantic field can be described systematically and separately. Furthermore, modules are also defined according to levels of frequency and distribution of lexemes in the

elexiko corpus. Right now, complex and comprehensive information on a module called "Dictionary on Public Discourse" is being compiled. It contains approximately 2,800 entries selected mainly according to their (high) frequency in the *elexiko* corpus.

Along with publishing the list of headwords (taken exclusively from the *elexiko* corpus) on the Internet in 2003, the *elexiko* dictionary was filled with sense-independent information for each headword generated automatically from the underlying corpus. This concerns 300,000 single-word entries comprising details on spelling, spelling variation, and syllabication. The orthographic information in particular was checked manually, because mistakes here are usually not tolerated by users. Since then, the project has been working on enriching as many entries as possible with further information generated automatically, *e.g.* automatically chosen citations (see Figure 1).

Choosing citations from the corpus is not carried out strictly according to rules of statistical concurrence, but by applying certain criteria, which help to improve the quality of the citations. For example, they have to be found in at least three different sources and come from at least three different years. Users may find these quotations helpful when they look up the meaning of a word. In addition to these text clippings, information on the coverage of the headword in the *elexiko* corpus is given. By showing the number of sources and years in which the headword occurs in the corpus, the user may get an idea of the distribution of the word.


Orthografie

Normgerechte Schreibung: Wörterbuch
Worttrennung: Wor|ter|buch

Belege (automatisch ausgewählt)

Die fünfte Klasse von Helfrich-Rall hat es inzwischen fast geschafft, auf ein einheitliches Niveau zu kommen, auch in Deutsch. "Jetzt hat die Reform richtigen Wettkampfcharakter bekommen", erzählt die Lehrerin lächelnd. Den Kleinen bereite es eine diebische Freude, sie beim Falschschreiben an der Tafel zu erwischen. Sowohl Helfrich-Rall als auch Vater mussen durchaus noch das **Worterbuch** bemühen. Denn selbst, "wenn wir uns bei ein paar Dingen wie dem Doppel-S problemlos umgestellt haben", meint Vater, "gibt es doch anderes, was der Gewöhnung bedarf." (M98/MAI.39667 Mannheimer Morgen, 12.05.1998, Ressort: Welt und Wissen; Kaum hat man alles kapiert, beginnt das Umlernen von neuem)

"ich furchte, daß mir hier meine Ehre genommen werden soll" könne "verwerflich" fur den Juristen etwas anderes sein als fur den Laien? nein - "verwerflich" bedeute auch und gerade für den Juristen ruchlos. Gnmms **Worterbuch** nennt als Beispiel den "verwerflichen Richter", der das Recht beugt. "und dieser Verwerflichkeit will man uns zeihen. (H95/FZ1.15914 Die Zeit, 25.01.1985, S. 06; Was heißt hier verwerflich?)

Er ist der Porsche unter den elektronischen Sprachencomputern: der neue Attaché von Hexaglot. Ausgestattet mit Sprachausgabe, SD-Card-Technologie, Lernsystem und Trainingsmodul führt die neueste Entwicklung der Langenscheidt-Tochter mit Sitz in der Sportallee 41 in zwei Sprachen (Deutsch, Englisch) sowie mit einem gastronomischen Spezialwortschatz in fünf Sprachen vorgewandt durch Reisen rund um den Globus. Insgesamt verfügt der Attaché über mehr als 5,1 Mio. Einträge. Mit Hilfe von SD-Cards kann das kleine Allround-Talent jederzeit um diverse **Worterbucher** und Wortschätze ergänzt werden. Preis: 279,90 Euro. (HMP07/MAR.01453 Hamburger Morgenpost, 13.03.2007, Beilage S. 7, Premiere für das Sprachgenie)

Dieses Stichwort gehört im *elexiko*-Korpus der Frequenzschicht VII (1.001-5.000 mal belegt) an. Es ist in 15 verschiedenen Zeitungen oder Zeitschriften aus 21 Jahrgängen belegt.

Weitere Informationen:
Automatisch ermitteltes Kookkurrenzprofil von **Worterbuch** in der CCDB
Grammatische Informationen (z.B. Angabe der Wortart, Flexionstabellen) unter canoo.net.

Figure 1. Automated information in elexiko on the headword "Wörterbuch"

Additionally in *elexiko*, there are hyperlinks to other online sources, where users may look up automatically compiled information on collocations (hyperlink to "Kookkurrenzdatenbank CCDB" developed at the IDS) and on grammar (hyperlink to canoo.net, where information on flexion and word formation is given). A direct link from dictionary entries to the underlying corpus has not yet been implemented, because many of the corpus texts are not openly accessible due to copyright. In the near future, *elexiko* will offer automatically compiled information on word formation with the headword. Words stemming from one headword will be given as hyperlinks, thus joining entries with lexicographically and automatically compiled information.

In *elexiko*, automated information is employed carefully; as much of this information as possible is checked manually in order to improve the quality. This has the negative effect of slowing down the process of publication and increasing the cost.

3. DWDS – Digital Dictionary of contemporary German

DWDS, a digital dictionary of contemporary German published at the Berlin-Brandenburg Academy of Science since 2004, was planned differently from the start (*cf.* Klein and Geyken 2000, 2001; Geyken 2005). This project aims at creating a "digital lexical system", that is easy to expand or correct and may be used for many different academic or non-academic purposes. DWDS combines a digitalised print dictionary with a word profile giving automated information on collocations, citations from an extensive corpus on German between 1900 and 1990, and a thesaurus. In the beta version of DWDS, which is shown here, on the first screen after looking up a word, all this information is combined (see Figure 2).

Figure 2. Information in DWDS on the headword "Wörterbuch"

The "Dictionary of Contemporary German" (WDG) was digitalised for DWDS, but was published in the 1960s and 70s in the German Democratic Republic and has been written on the basis of a paper archive of citations. For the online version, the possibility of presenting only part of the information in the entry has been implemented, and information on pronunciation will soon be added.

The thesaurus incorporated is not developed by DWDS, but OpenThesaurus maintains its own domain, where anybody may contribute to the dictionary. Its information is built into the DWDS site, but not hyperlinked with other information. Synonyms given, for example, are not hyperlinked to entries in the WDG dictionary.

The citations quoted come from the DWDS corpora and are chosen automatically. Usually 10 KWICs are given, but full contexts and more KWICs may be opened. Although the DWDS corpora were planned carefully, the quality of the citations given is not always convincing, as with any automatic selection. But even more important is that the DWDS corpora were not the basis for the WDG dictionary.

The word profile gives words and phrases collocating with the headword in the DWDS corpora in a word cloud. The word cloud shows the most frequent words co-occurring with the headword, but in many cases those words are not part of the WDG dictionary entry itself and vice versa. This is of course the case because the DWDS corpora are not the basis for the WDG dictionary. Here, automated information from one source and lexicographically written information from another source do not really harmonise, but at least they complement each other.

4. Conclusion

Two online reference works for German approach the matter of incorporating automated and lexicographically compiled information on words completely differently. While DWDS has compiled a lot of information on German from different sources in quite a short time and presents it in one user interface without tagging each kind of information, *elexiko* has less information, but hyperlinks to other applications. Automatically compiled information is checked lexicographically in *elexiko* as much as possible, slowing down the process of publication. Automated information is also labelled as such. In addition to this, new corpus-based, complex and comprehensive information on very frequent entries is being compiled in *elexiko*.

When contrasting information on paradigmatic relations in DWDS and *elexiko*, the huge difference in quality and quantity between automated information and lexicographically compiled information becomes apparent. The OpenThesaurus in DWDS gives the following synonyms (single words and multi-word units) for the headword *Aids* (http://beta.dwds.de/?qu=Aids&view=1): *Acquired Immune Deficiency Syndrome*, *AIDS*, *erworbenes Immunschwäche-Syndrom* (*i.e.* 'acquired immune deficiency syndrome'), and it records *Infektionskrankheit* (*i.e.* 'infectious disease') as a hypernym. In its word profile, DWDS names the collocates *Armut* (*i.e.* 'poverty'), *Malaria* (*i.e.* 'malaria'), *Tuberkulose* (*i.e.* 'tuberculosis') and *sterben an* (*i.e.* 'to die

of'). Three of these collocates would probably be classified as paradigmatic partners in *elexiko*, the verbal phrase *an Aids sterben* (*i.e.* 'to die of Aids') would appear as typical usage.

In the *elexiko* entry for *Aids* (http://www.owid.de/pls/db/p4_anzeige.lesart?v_id=302141&v_lesart=Krankheit) synonyms given are *Immunschwäche* (*i.e.* 'immune deficiency') and *Immunschwächekrankheit* (*i.e.* 'immune deficiency disease'), hypernyms are *Epidemie* (*i.e.* 'epidemic'), *Erkrankung* (*i.e.* 'disease'), *Krankheit* (*i.e.* 'illness'), *Infektionskrankheit* (*i.e.* 'infectious disease') and *Seuche* (*i.e.* 'epidemic'). There are also three thematically defined groups of incompatible (*i.e.* cohyponym) partner words: *Armut* (*i.e.* 'poverty') and *Hunger* (*i.e.* 'hunger'), *Geschlechtskrankheit* (*i.e.* 'veneral disease') and *HIV-Infektion* (*i.e.* 'HIV infection'), *Alkohol* (*i.e.* 'alcohol') and *Droge* (*i.e.* 'drug'). Each paradigmatic partner is accompanied by a citation illustrating the relation (*cf.* Storjohann 2005a). In addition, information on each type of paradigmatic relation can be opened.

It is not for lexicographers to decide which way is to be preferred, but for the users. We do not yet know whether users would like to be able to rate the reliability of lexical information in online dictionaries. We do not even know how users will respond to automated information in a dictionary in general. Will users appreciate the comprehensive description of paradigmatic relations in *elexiko* or will automated information as in DWDS satisfy them in specific instances of dictionary use? Lexicographers can only assess the possibilities computational linguistics offers, and test new ways of enriching electronic dictionaries with lexicographically compiled and automated information, linking them in a fruitful way. Only extensive usage research, as intended for the *elexiko* project, will help to find answers to these questions.

References

AGIRRE, E. and EDMONDS, P. (eds). (2006). *Word sense disambiguation: Algorithms and applications.* Dordrecht: Springer.

BAUGH, S., HARLEY, A. and JELLIS, S. (1996). The Role of Corpora in Compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics* 1/1: 39-59.

BOGURAEV, B.K. (1993). The contribution of computational lexicography. In M. Bates and R.M. Weischedel (eds). *Challenges in Natural Language Processing*. Cambridge: Cambridge University Press: 99-134.

DWDS – Digital Dictionary of the German Language of the 20th Century: http://www.dwds.de/ (02.11.2009).

elexiko: http://www.owid.de/elexiko_/index.html (02.11.2009).

Euralex Bibliography of Lexicography: http://euralex.pbwiki.com/ (02.11.2009).

GEYKEN, A. (2005). Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS). *BBAW Circular,* 32: 40.

HAß, U. (2005). *Grundfragen der elektronischen Lexikographie. elexiko – Das Online-Informationssystem zum deutschen Wortschatz.* Berlin/New York: de Gruyter.

KLEIN, W. and GEYKEN, A. (2000). Projekt «Digitales Wörterbuch der deutschen Sprache des 20. Jh.». *Jahrbuch der BBAW*, 1999: 277-289.

KLEIN, W. and GEYKEN, A. (2001). Projekt «Digitales Wörterbuch der deutschen Sprache des 20. Jh. ». *Jahrbuch der BBAW*, 2000: 263-270.

KLOSA, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In W. Kallmeyer and G. Zifonun (eds). *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*. Berlin/New York: de Gruyter: 105-122.

KLOSA, A., SCHNÖRCH, U. and STORJOHANN, P. (2006). ELEXIKO – A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim. In C. Marello *et al.* (eds). *Proceedings of the 12th EURALEX International Congress* (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th-9th, 2006. Vol. 1. Turin: Edizioni dell'Orso Alessandria: 425-430.

MÖNNICH, U. and KÜHNBERGER, K.-U. (eds). (2008). *Foundations of Ontologies in Text Technology, Part II: Applications*. (*Zeitschrift für Computerlinguistik und Sprachtechnologie*, 23/1).

OBELEX – Online Bibliography of Electronic Lexicography: http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht (02.11.2009).

SINCLAIR, J. (1987). *Looking Up. An Account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London: Harper Collins.

STORJOHANN, P. (2005a). Corpus-driven *vs.* corpus-based approach to the study of relational patterns. In *Proceedings of the Corpus Linguistics Conference 2005 in Birmingham. Vol. 1, no. 1.* (http://www.corpus.bham.ac.uk/PCLC/).

STORJOHANN, P. (2005b). *elexiko* – A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics,* 34: 55-83.

TEUBERT, W. (1999). Korpuslinguistik und Lexikographie. *Deutsche Sprache,* 4: 292-313.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: Benjamins.